A Systematic Analysis of Translation Model Search Spaces

A Systematic Analysis of Translation Model Search Spaces

Michael Auli, Adam Lopez, Hieu Hoang and **Philipp Koehn**
University of Edinburgh
10 Crichton Street
Edinburgh, EH8 9AB
United Kingdom
m.auli@sms.ed.ac.uk, alopez@inf.ed.ac.uk, h.hoang@sms.ed.ac.uk, pkoehn@inf.ed.ac.uk

Abstract

Translation systems are complex, and most metrics do little to pinpoint causes of error or isolate system differences. We use a simple technique to discover induction errors, which occur when good translations are absent from model search spaces. Our results show that a common pruning heuristic drastically increases induction error, and also strongly suggest that the search spaces of phrase-based and hierarchical phrase-based models are highly overlapping despite the well known structural differences.

1 Introduction

Most empirical work in translation analyzes models and algorithms using BLEU (Papineni et al., 2002) and related metrics. Though such metrics are useful as sanity checks in iterative system development, they are less useful as analytical tools. The performance of a translation system depends on the complex interaction of several different components. Since metrics assess only output, they fail to inform us about the consequences of these interactions, and thus provide no insight into the errors made by a system, or into the design tradeoffs of competing systems.

In this work, we show that it is possible to obtain such insights by analyzing translation system components in isolation. We focus on model search spaces (§2), posing a very simple question: *Given a model and a sentence pair, does the search space contain the sentence pair?* Applying this method to the analysis and comparison of French-English translation using both phrase-based and hierarchical phrase-based systems yields surprising results, which we analyze quantitatively and qualitatively.

- First, we analyze the **induction error** of a model, a measure on the completeness of the search space. We find that low weight phrase translations typically discarded by heuristic pruning nearly triples the number of reference sentences that can be exactly reconstructed by either model (§3).
- Second, we find that the high-probability regions in the search spaces of phrase-based and hierarchical systems are nearly identical (§4). This means that reported differences between the models are due to their rankings of competing hypotheses, rather than structural differences of the derivations they produce.

2 Models, Search Spaces, and Errors

A translation model consists of two distinct elements: an unweighted ruleset, and a parameterization (Lopez, 2008a; 2009). A **ruleset** licenses the steps by which a source string $f_1 ... f_I$ may be rewritten as a target string $e_1 ... e_J$. A **parameterization** defines a weight function over every sequence of rule applications.

In a phrase-based model, the ruleset is simply the unweighted phrase table, where each phrase pair $f_i ... f_{i'} / e_j ... e_{j'}$ states that phrase $f_i ... f_{i'}$ in the source can be rewritten as $e_j ... e_{j'}$ in the target. The model operates by iteratively applying rewrites to the source sentence until each source word has been consumed by exactly one rule. There are two additional heuristic rules: The distortion limit dl constrains distances over which phrases can be reordered, and the translation option limit tol constrains the number of target phrases that may be considered for any given source phrase. Together, these rules completely determine the finite set of all possible target sentences for a given source sentence. We call this set of target sentences the **model search space**.

The parameterization of the model includes all information needed to score any particular se-

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 224–232,
Athens, Greece, 30 March – 31 March 2009.

quence of rule applications. In our phrase-based model, it typically includes phrase translation probabilities, lexical translation probabilities, language model probabilities, word counts, and coefficients on the linear combination of these. The combination of large rulesets and complex parameterizations typically makes search intractable, requiring the use of **approximate search**. It is important to note that, regardless of the parameterization or search used, the set of all possible output sentences is still a function of *only* the ruleset.

Germann et al. (2004) identify two types of translation system error: **model error** and **search error**.[1] Model error occurs when the optimal path through the search space leads to an incorrect translation. Search error occurs when the approximate search technique causes the decoder to select a translation other than the optimum.

Given the decomposition outlined above, it seems clear that model error depends on parameterization, while search error depends on approximate search. However, there is no error type that clearly depends on the ruleset (Table 1). We therefore identify a new type of error on the ruleset: **induction error**. Induction error occurs when the search space does not contain the correct target sentence at all, and is thus a more fundamental defect than model error. This is difficult to measure, since there could be many correct translations and there is no way to see whether they are all absent from the search space.[2] However, if we assume that a given reference sentence is ground truth, then as a proxy we can simply ask whether or not the model search space contains the reference. This assumption is of course too strong, but over a sufficiently large test set, it should correlate with metrics which depend on the reference, since under most metrics, exactly reproducing the reference results in a perfect score. More loosely, it should correlate with translation accuracy—even if there are many good translations, a model which is systematically unable to produce any reference sentences from a sufficiently large test sample is almost certainly deficient in some way.

3 Does Ruleset Pruning Matter?

The heuristic translation option limit tol controls the number of translation rules considered per source span. It plays a major role in keeping the search space manageable. Ignoring reordering, the complexity of the search in a phrase-based model is $O(n^{tol})$, where n is the number of French spans. Therefore tol has a major effect on efficiency. Tight pruning with tol is often assumed without question to be a worthwhile tradeoff. However, we wish to examine this assumption more closely.

Consider the French word *problème*. It has 288 different translation options in the phrase table of our French-English phrase-based system. The phrase translation probability $p(e|f)$ over these options is a familiar Zipf distribution (Figure 1). The most likely candidate translation for the word is *problem* with a probability of 0.71, followed by *issue* with a much smaller probability of 0.12. Further down, we find *challenge* at rank 25, *obstacle* at 44 and *dilemma* at rank 105. Depending on the context, these might be perfectly good translations. However, with a typical tol of 20, most of these options are not considered during decoding.

Table 2 shows that 93.8% of rules are available during decoding with the standard tol setting and only about 0.1% of French spans of the entire ruleset have more than 20 translation options. It seems as if already most of the information is available when using the default limit. However, a tol of 20 can clearly exclude good translations as illustrated by our example. Therefore we hypothesize the following: *Increasing the translation option limit gives the decoder a larger vocabulary which in turn will decrease the induction error*. We sup-

ruleset	**induction error**
parameterization	model error
search	search error

Table 1: Translation system components and their associated error types.

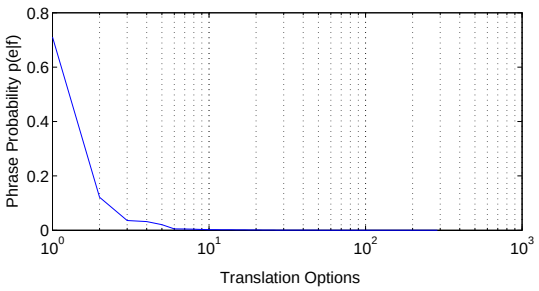

Figure 1: Distribution $p(f|e)$ of the English translation options for the French word *problème*.

[1] They also identify variants within these types.

[2] It can also be gamed by using a model that can generate any English word from any French word. However, this is not a problem for the real models we investigate here.

tol	Ruleset Size	French Spans
20	93.8	99.9
50	96.8	100.0
100	98.3	100.0
200	99.2	100.0
400	99.7	100.0
800	99.9	100.0
All	100.0	100.0

Table 2: Ruleset size expressed as percentage of available rules when varying the limit of translation options tol per English span and percentage of French spans with up to tol translations.

port this hypothesis experimentally in §5.4.

4 How Similar are Model Search Spaces?

Most work on hierarchical phrase-based translation focuses quite intently on its structural differences from phrase-based translation.

- A hierarchical model can translate discontiguous groups of words as a unit. A phrase-based model cannot. Lopez (2008b) gives indirect experimental evidence that this difference affects performance.
- A standard phrase-based model can reorder phrases arbitrarily within the distortion limit, while the hierarchical model requires some lexical evidence for movement, resorting to monotone translation otherwise.
- While both models can indirectly model word deletion in the context of phrases, the hierarchical model can delete words using non-local context due to its use of discontiguous phrases.

The underlying assumption in most discussions of these models is that these differences in their generative stories are responsible for differences in performance. We believe that this assumption should be investigated empirically.

In an interesting analysis of phrase-based and hierarchical translation, Zollmann et al. (2008) forced a phrase-based system to produce the translations generated by a hierarchical system. Unfortunately, their analysis is incomplete; they do not perform the analysis in both directions. In §5.5 we extend their work by requiring each system to generate the 1-best output of the other. This allows us to see how their search spaces differ.

5 Experiments

We analyse rulesets in isolation, removing the influence of the parametrization and heuristics as much as possible for each system as follows: First, we disabled beam search to avoid pruning based on parametrization weights. Second, we require our decoders to generate the reference via disallowing reference-incompatible hypothesis or chart entries. This leaves only some search restrictions such as the distortion limit for the phrase-based system for which we controlled, or the maximum number of source words involved in a rule application for the hierarchical system.

5.1 Experimental Systems

Our phrase-based system is Moses (Koehn et al., 2007). We set its stack size to 10^5, disabled the beam threshold, and varied the translation option limit tol. Forced translation was implemented by Schwartz (2008) who ensures that hypothesis are a prefix of the reference to be generated.

Our hierarchical system is Hiero (Chiang, 2007), modified to construct rules from a small sample of occurrences of each source phrase in training as described by Lopez (2008b). The search parameters restricting the number of rules or chart entries as well as the minimum threshold were set to very high values (10^{50}) to prevent pruning. Forced translation was implemented by discarding rules and chart entries which do not match the reference.

5.2 Experimental Data

We conducted experiments in French-English translation, attempting to make the experimental conditions for both systems as equal as possible. Each system was trained on French-English Europarl (Koehn, 2005), version 3 (40M words). The corpus was aligned with GIZA++ (Och and Ney, 2003) and symmetrized with the grow-diag-final-and heuristic (Koehn et al., 2003). A trigram language model with modified Kneser-Ney discounting and interpolation was used as produced by the SRILM toolkit (Stolcke, 2002). Systems were optimized on the WMT08 French-English development data (2000 sentences) using minimum error rate training (Och, 2003) and tested on the WMT08 test data (2000 sentences). Rules based on unaligned words at the edges of foreign and source spans were not allowed unless otherwise stated, this is denoted as the *tightness con-*

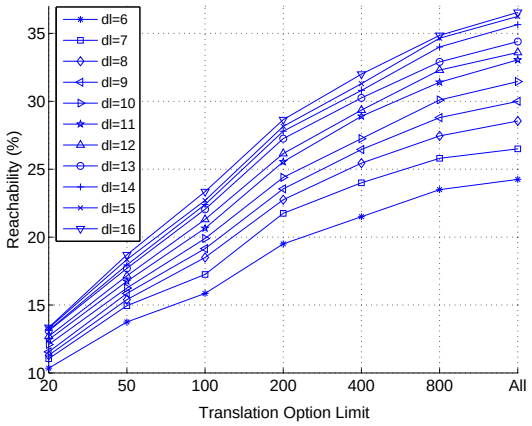

Figure 2: Coverage for phrase-based reference aligned translation on test data when varying the translation option and the distortion limits (dl).

French Spans	Number of Translations
des	3006
les	2464
la	1582
de	1557
en	1428
de la	1332
fait	1308
une	1303
à	1291
le	1273
d’	1271
faire	1263
l’	1111
c’ est	1109
à la	1053
,	1035

Table 3: French spans with more than 1000 translation options.

straint. Ayan and Dorr (2006) showed that under certain conditions, this constraint could have significant impact on system performance. The maximum phrase lengths for both the hierarchical and phrase-based system were set to 7. The distortion limit (dl) for the phrase-based system was set to 6 unless otherwise mentioned. All other settings were left at their default values as described by Chiang (2007) and Koehn et al. (2007).

5.3 Metric: Reference Reachability

We measure system performance in terms of **reference reachability**, which is the inverse of induction error: A system is required to be able to exactly reproduce the reference, otherwise we regard the result as an error.

5.4 Analysis of Ruleset Pruning

In §3 we outlined the hypothesis that increasing the number of English translation options per French span can increase performance. Here we present results for both phrase-based and hierarchical systems to support this claim.

5.4.1 Quantitative Results

Figure 2 shows the experimental results when forcing our phrase-based system to generate unseen test data. We observe more than 30% increase in reachability from $tol = 20$ to $tol = 50$ for all $dl \geq 6$ which supports our hypothesis that increasing tol by a small multiple can have a significant impact on performance. With no limit on tol, reachability nearly triples.

Notably, the increase stems from the small fraction of French spans (0.1%) which have more than 20 translation options (Table 2). There are only 16 French spans (Table 3) which have more than 1000 translation options, however, utilising these can still achieve an increase in reachability of up to 5%. The list shown in Table 3 includes common articles, interpuncutation, conjunctions, prepositions but also verbs which have unreliable alignment points and therefore a very long tail of low probability translation options. Yet, the largest increase does not stem from using such unreliable translation options, but rather when increasing tol by a relatively small amount.

The increases we see in reachability are proportional to the size of the ruleset: The highest increases in ruleset size can be seen between $tol = 20$ and $tol = 200$ (Table 2), similarly, reachability performance has then the largest increase. For higher tol settings both the increases of ruleset size and reachability are smaller.

Figure 3 plots the average number of words per sentence for the reachable sentences. The average sentence length increases by up to six words when using all translation options. The black line represents the average number of words per sentence of the reference set. This shows that longer and more complex sentences can be generated when using more translation options.

Similarly, for our hierarchical system (see Fig-

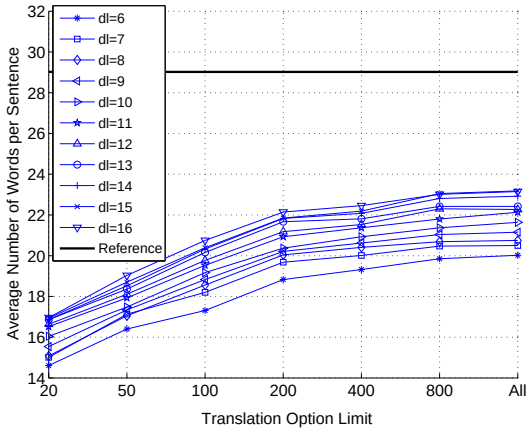

Figure 3: Average number of words per sentence for the reachable test data translations of the phrase-based system (as shown in Figure 2).

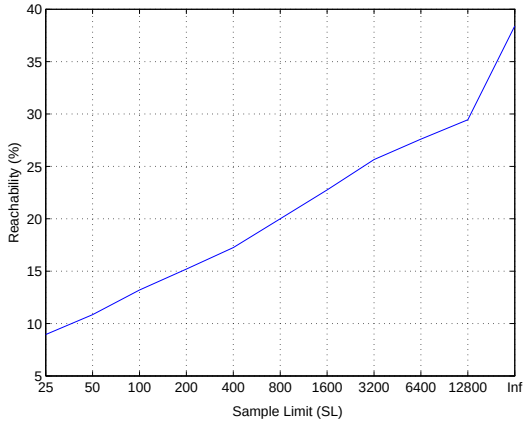

Figure 4: Coverage for hierarchical reference aligned translation on test data when varying the number of matching French samples (sl) drawn from the training data. The baseline setting is $sl = 300$.

ure 4) we find that reachability can be more than doubled when drawing a richer ruleset sample than in the baseline setting. Those results are not directly comparable to the phrase-based system due to the slightly different nature of the parameters which were varied: In the phrase-based case we have tol different English spans per French span. In the hierarchical system it is very likely to have duplicate French spans in the sample drawn from training data. Yet, the trend is the same and thus supports our claim.

5.4.2 Qualitative Results

We were interested how the performance increase could be achieved and therefore looked into which kind of translation options were involved when a translation was generable with a higher tol setting. One possibility is that the long tail of translation options includes all kinds of English spans that match some part of the reference but are simply an artifact of unreliable alignment points.

We looked at the first twenty translations produced by our phrase-based system under $dl = 10$ which could not be generated with $tol = 20$ but with $tol = 50$. The aim was to find out which translation options made it possible to reach the reference under $tol = 50$.

We found that nearly half (9) involved translation options which used a common or less common translation of the foreign span. The first four translations in Table 4 are examples for that. When allowing unaligned words at the rule edges it turns out that even 13 out of 20 translations are based on sound translation options.

The remaining sentences involved translation options which were an artifact of unreliable alignment points. An example rule is *la* / *their*, which erroneously translates a common determiner into an equally common adjective. The last translation in Figure 4 involves such a translation option.

This analysis demonstrates that the performance increase between $tol = 20$ to $tol = 50$ is to a considerable extent based on translation options which are meaningful.

5.5 Analysis of Mutual Reachability

The aim of this analysis was to find out by how much the high-probability search spaces of the phrase-based and hierarchical models differ. The necessary data was obtained via forcing each system to produce the 1-best translation of the other system denoted as the *unconstrained translation*. This unconstrained translation used the standard setting for the number of translation options.

We controlled for the way unaligned words were handled during rule extraction: The phrase-based system allowed unaligned words at the edges of phrases while the hierarchical system did not. We varied this condition for the phrase-based system. The distortion limit of the phrase-based system was set to 10. This is equal to the maximum span a rule can be applied within the hierarchical system.

We carried out the same experiment for German-English and English-German translation which serve as examples for translating into a mor-

S:	je voterai en **faveur** du projet de règlement .
R:	i will vote to **approve** the draft regulation .
O:	i shall be voting in favour of the draft regulation .
S:	... il npeut y avoir de délai transitoire **en matière de** respect des règles démocratiques .
R:	... there can be no transitional period **for** complying with democratic rules .
O:	... there can be no transitional period in the field of democratic rules .
S:	je souhaite aux négociateurs la **poursuite du** succès de leur travail dans ce domaine important .
R:	i wish the negotiators **continued** success with their work in this important area .
O:	i wish the negotiators the continuation of the success of their work on this important area .
S:	mais commencons par les **points positifs** .
R:	but let us begin with the **good news** .
O:	but let us begin with the positive points .
S:	... partage la plupart des conclusions que tire **le rapporteur** .
R:	... share the majority of conclusions that **he** draws .
O:	... share most of the conclusions that is the rapporteur .

Table 4: Example translations which could be generated with $tol = 50$ but not with $tol = 20$. For each translation the source (S), reference (R) and the unconstrained output (O) are shown. Bold phrases mark translation options which were not available under $tol = 20$.

phologically simpler and more complex language respectively. The test and training sets for these languages are similarly sized and are from the WMT08 shared task.

5.5.1 Quantitative Results

Table 5 shows the mutual reachability performance for our phrase-based and hierarchical system. The hierarchical system can generate almost all of the 1-best phrase-based translations, particularly when unaligned words at rule edges are disallowed which is the most equal condition we experimented with. The phrase-based reachability for English-German using tight rulesets is remarkably low. We found that this is because the hierarchical model allows unaligned words around gaps under the tight constraint. This makes it very hard for the phrase-based system to reach the hierarchical translation. However, the phrase-based system can overcome this problem when the tightness constraint is loosened (last row in Table 5).

Table 6 shows the translation performance measured in BLEU for both systems for normal unconstrained translation. It can be seen that the difference is rather marginal which is in line with our reachability results.

We were interested why certain translations of one system were not reachable by the other system. The following two subsections describe our analysis of these translations for the French-English language pair.

Translation Direction	fr-en	de-en	en-de
$H_t \rightarrow P_t$	99.40	97.65	98.50
$H_t \rightarrow P_{nt}$	95.95	93.95	94.30
$P_t \rightarrow H_t$	93.75	92.30	82.95
$P_{nt} \rightarrow H_t$	97.55	97.55	96.30

Table 5: Mutual reachability performance for French-English (fr-en), German-English (de-en) and Enlgish-German (en-de). P $\rightarrow$ H denotes how many hierarchical (H) high scoring outputs can be reached by the phrase-based (P) system. The subscripts nt (non-tight) and t (tight) denote the use of rules with unaligned words or not.

5.5.2 Qualitative Analysis of Unreachable Hierarchical Translations

We analysed the first twenty translations within the set of unreachable hierarchical translations when disallowing unaligned words at rule edges to find out why the phrase-based system fails to reach them. Two aspects were considered in this analysis: First, the successful hierarchical derivation and second, the relevant part of the phrase-based ruleset which was involved in the failed forced translation i.e. how much of the input and the reference could be covered by the raw phrase-pairs available to the phrase-based system.

Within the examined subset, the majority of sentences (14) involved hierarchical rules which could not be replicated by the phrase-based sys-

System	fr-en	de-en	en-de
Phrase-based	31.96	26.94	19.96
Hierarchical	31.62	27.18	20.20
Difference absolute	0.34	0.24	0.24
Difference (%)	1.06	0.90	1.20

Table 6: Performance for phrase-based and hierarchical systems in BLEU for French-English (fr-en), German-English (de-en) and English-German (en-de).

tem. We described this as the first structural difference in §4. Almost all of these translations (12 out of 14) could not be generated because of the third structural difference which involved rule that omits the translation of a word within the French span. An example is the rule $X \rightarrow$ *est*$X_{\boxed{1}}$*ordinaire*$X_{\boxed{2}}$/*is*$X_{\boxed{1}}X_{\boxed{2}}$ which omits a translation for the French word *ordinaire* in the English span. For this particular subset the capability of the hierarchical system to capture long-distance reorderings did not make the difference, but rather the ability to drop words within a translation rule.

The phrase-based system cannot learn many rules which omit the translation of words because we disallowed unaligned words at phrase edges. The hierarchical system has the same restriction, but the constraint does not prohibit rules which have unaligned words *within* the rule. This allows the hierarchical system to learn rules such as the one presented above. The phrase-based system can learn similar knowledge, although less general, if it is allowed to have unaligned words at the phrase edges. In fact, without this constraint 13 out of the 20 analysed rules can be generated by the phrase-based system.

Figure 5 shows a seemingly simple hierarchical translation which fails to be constructed by the phrase-based system: The second rule application involves both the reordering of the translation of *postaux* and the omittance of a translation for *concurrence*. This translation could be easily captured by a phrase-pair, however, it requires that the training data contains exactly such an example which was not the case. The closest rule the phrase-based rulestore contains is *des services postaux* / *postal services* which fails since it does not cover all of the input. This is an example for when the generalisation of the hierarchical model is superior to the phrase-based approach.

5.5.3 Qualitative Analysis of Unreachable Phrase-based Translations

The size of the set of unreachable phrase-based translations is only 0.6% or 12 sentences. This means that almost all of the 1-best outputs of the phrase-based translations can be reached by the hierarchical system. Similarly to above, we analysed which words of the input as well as which words of the phrase-based translation can be covered by the available hierarchical translation rules.

We found that all of the translations were not generable because of the second structural difference we identified in §4. The hierarchical ruleset did not contain a rule with the necessary lexical evidence to perform the same *reordering* as the phrase-based model. Figure 6 shows a phrase-based translation which could not be reached by the hierarchical system because a rule of the form $X \rightarrow$ *électorales*$X_{\boxed{1}}$/$X_{\boxed{1}}$*electoral* would be required to move the translation of *électorales* (electoral) just before the translation of *réunions* (meetings). Inspection of the hierarchical ruleset reveals that such a rule is not available and so the translation cannot be generated.

The small size of the set of unreachable phrase-based translations shows that the lexically informed reordering mechanism of the hierarchical model is not a large obstacle in generating most of the phrase-based outputs.

In summary, each system can reproduce nearly all of the highest-scoring outputs of the other system. This shows that the 1-best regions of both systems are nearly identical despite the differences discussed in §4. This means that differences in observed system performance are probably attributable to the degree of model error and search error in each system.

6 Related Work and Open Questions

Zhang et al. (2008) and Wellington et al. (2006) answer the question: what is the minimal grammar that can be induced to completely describe a training set? We look at the related question of what a heuristically induced ruleset can translate in an unseen test set, considering both phrase- and grammar-based models. We also extend the work of Zollmann et al. (2008) on Chinese-English, performing the analysis in both directions and providing a detailed qualitative explanation.

Our focus has been on the induction error of models, a previously unstudied cause of transla-

```
Source:       concurrence des services postaux
Reference:    competition between postal services
Hierarchical: postal services
Deviation:
( [0-4: @S  ->  @X^1  |  @X^1 ]
     ( [0-4: @X  ->  concurrence @X^1 postaux  |  postal @X^1 ]       postal
          ( [1-3: @X  ->  des services  |  services ]      services
          )
     )
)
```

Figure 5: Derivation of a hierarchical translation which cannot be generated by the phrase-based system, in the format of Zollmann et al. (2008). The parse tree contains the outputs (shaded) at its leaves in infix order and each non-leaf node denotes a rule, in the form: [Source-span: LHS →RHS].

```
Source:       ceux qui me disaient cela faisaient par exemple rèfèrence à certaines des
              réunions électorales auxquelles ils avaient assisté .
Phrase-based: those who said to me that were for example refer to some of which
              they had been electoral meetings .
Reference:    they referred to some of the election meetings , for example , that
              they had gone to .
```

Figure 6: Phrase-based translation which cannot be reached by the hierarchical system because no rule to perform the necessary reordering is available. Marked sections are source and reference spans involved in the largest possible partial hierarchical derivation.

tion errors. Although the results described here are striking, our exact match criterion for reachability is surely too strict—for example, we report an error if even a single comma is missing. One solution is to use a more tolerant criterion such as WER and measure the amount of deviation from the reference. We could also maximize BLEU with respect to the reference as in Dreyer et al. (2007), but it is less interpretable.

7 Conclusion and Future Work

Sparse distributions are common in natural language processing, and machine translation is no exception. We showed that utilizing more of the entire distribution can dramatically improve the coverage of translation models, and possibly their accuracy. Accounting for sparsity explicitly has achieved significant improvements in other areas such as in part of speech tagging (Goldwater and Griffiths, 2007). Considering the entire tail is challenging, since the search space grows exponentially with the number of translation options. A first step might be to use features that facilitate more variety in the top 20 translation options. A more elaborate aim is to look into alternatives to maximum likelihood hood estimation such as in Blunsom and Osborne (2008).

Additionally, our expressiveness analysis shows clearly that the 1-best region of hierarchical and phrase-based models is nearly identical. Discounting cases in which systems handle unaligned words differently, we observe an overlap of between 96% and 99% across three language pairs. This implies that the main difference between the models is in their parameterization, rather than in the structural differences in the types of translations they can produce. Our results also suggest that the search spaces of both models are highly overlapping: The results for the 1-best region allow the conjecture that also other parts of the search space are behaving similarly since it appears rather unlikely that spaces are nearly disjoint with only the 1-best region being nearly identical. In future work we aim to use n-best lists or lattices to more precisely measure search space overlap. We also aim to analyse the effects of the model and search errors for these systems.

Acknowledgements

This research was supported by the Euromatrix Project funded by the European Commission (6th Framework Programme). The experiments were conducted using the resources provided by the Edinburgh Compute and Data Facility (ECDF). Many thanks to the three anonymous reviewers for very helpful comments on earlier drafts.

References

N. F. Ayan and B. Dorr. 2006. Going beyond AER: An extensive analysis of word alignments and their impact on MT. In *Proc. of ACL-COLING*, pages 9–16, Jul.

P. Blunsom and M. Osborne. 2008. Probabilistic inference for machine translation. In *Proc. of EMNLP*.

D. Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

M. Dreyer, K. B. Hall, and S. P. Khudanpur. 2007. Comparing reordering constraints for SMT using efficient BLEU oracle computation. In *Proc. of Workshop on Syntax and Structure in Statistical Translation*, pages 103–110, Apr.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. 2004. Fast and optimal decoding for machine translation. *Artificial Intelligence*, 154(1–2):127–143, Apr.

S. Goldwater and T. Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proc. of ACL*, pages 744–751, Prague, Czech Republic, June.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of HLT-NAACL*, pages 48–54, Morristown, NJ, USA.

P. Koehn, H. Hoang, A. B. Mayne, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL Demonstration Session*, pages 177–180, Jun.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

A. Lopez. 2008a. Statistical machine translation. *ACM Computing Surveys*, 40(3).

A. Lopez. 2008b. Tera-scale translation models via pattern matching. In *Proc. of COLING*, pages 505–512, Aug.

A. Lopez. 2009. Translation as weighted deduction. In *Proc. of EACL*.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, pages 160–167, Morristown, NJ, USA.

K. Papineni, S. Roukos, T. Ward, and W. jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.

L. Schwartz. 2008. Multi-source translation methods. In *Proc. of AMTA*, October.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Int. Conf. Spoken Language Processing (ICSLP 2002)*.

B. Wellington, S. Waxmonsky, and I. D. Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proc. of ACL*, pages 977–984, Morristown, NJ, USA.

H. Zhang, D. Gildea, and D. Chiang. 2008. Extracting synchronous grammar rules from word-level alignments in linear time. In *Proc. of COLING*, pages 1081–1088, Manchester, UK.

A. Zollmann, A. Venugopal, F. Och, and J. Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proc. of COLING*.